DOCUMENT RESUME

ED 454 862 IR 058 153

AUTHOR	Chan, Lois Mai
TITLE	Exploiting LCSH, LCC, and DDC To Retrieve Networked Resources: Issues and Challenges.
PUB DATE	2000-11-00
NOTE	21p.; In: Bicentennial Conference on Bibliographic Control for the New Millennium: Confronting the Challenges of Networked Resources and the Web (Washington, DC, November 15-17, 2000); see IR 058 144.
AVAILABLE FROM	For full text: http://lcweb.loc.gov/catdir/bibcontrol/chan_paper.html.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Access to Information; *Cataloging; Dewey Decimal Classification; *Indexing; *Information Retrieval; Library of Congress Classification; Models; Subject Index Terms; *World Wide Web
IDENTIFIERS	*Electronic Resources; Library of Congress Subject Headings

ABSTRACT

This paper examines how the nature of the World Wide Web and characteristics of networked resources affect subject access and analyzes the requirements of effective indexing and retrieval tools. The current and potential uses of existing tools and possible courses of future development are explored in the context of recent research. The first section addresses the new environment, including the nature of the online public access catalog (OPAC), characteristics of traditional library tools, and differences between electronic resources and traditional library materials. The second section discusses retrieval models, including the Boolean, vector, and probabilistic models. The third section covers subject access on the Web, including functional requirements of subject access tools, operational requirements, verbal subject access, and classification/subject categorization. The fourth section describes recent research on subject access systems, including automatic indexing, mapping terms and data from different sources, and integrating different subject access tools. The fifth section examines traditional tools in the networked environment, including Library of Congress Subject Headings (LCSH), Library of Congress Classification (LCC), and Dewey Decimal Classification (DDC). (Contains 48 references.) (MES)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 454 862

Exploiting LCSH, LCC, and DDC To Retrieve Networked Resources Issues and Challenges

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

B. Wiggins

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Lois Mai Chan
School of Library and Information Science
University of Kentucky
Lexington, KY 40506-0039

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Final version

Introduction

The proliferation and the infinite variety of networked resources and their continuing rapid growth present enormous opportunities as well as unprecedented challenges to library and information professionals. The need to integrate Web resources with traditional types of library materials has necessitated a re-examination of the established, well-proven tools that have been used in bibliographic control. Librarians confront new challenges in extending their practice in selecting and organizing library materials to a variety of resources in a dynamic networked environment. In this environment, the tension between quality and quantity has never been keener. Providing quality access to a large quantity of resources poses special challenges.

This paper examines how the nature of the Web and characteristics of networked resources affect subject access and analyses the requirements of effective indexing and retrieval tools. The current and potential uses of existing tools and possible courses for future development will be explored in the context of recent research.

A New Environment and Landscape

For centuries librarians have addressed issues of information storage and retrieval and have developed tools that are effective in handling traditional materials. However, any deliberation on the future of traditional tools should take into consideration the characteristics of networked resources and the nature of information retrieval on the Web. The sheer size demands efficient tools; it is a matter of economy. I will begin by reviewing briefly the nature of the OPAC and the characteristics of traditional library resources. OPACs are by and large homogeneous, at least in terms of content organization and format of presentation, if not in interface design. They are standardized due to the common tools

IR058153

BEST COPY AVAILABLE

(ÀÅCR2R/MARC, LCSH, LCC, DDC, etc.) used in their construction, and there is a level of consistency among them. The majority of resources represented in the OPACs, i.e., traditional library materials, typically manifest the following characteristics:

- tangible (they represent physical items)
- well-defined (they can be defined and categorized in terms of specific types, such as books, journals, maps, sound recordings, etc.)
- self-contained (they are packaged in recognizable units)
- relatively stable (though subject to physical deterioration, they are not volatile)

The World Wide Web, on the other hand, can be described as vast, distributed, multifarious, machine-driven, dynamic/fluid, and rapidly evolving. Electronic resources, in contrast to traditional library materials, are often:

- amorphous
- ill-defined
- not self-contained
- unstable
- volatile

Over the years, standards and procedures for organizing library materials have been developed and tested. Among these is the convention that trained catalogers and indexers typically carry the full responsibility for providing metadata through cataloging and indexing. In contrast, the networked environment is still developing, meaning that appropriate and efficient methods for resources description and organization are still evolving. Because of the sheer volume of electronic resources, many people without formal training in bibliographic control, including subject specialists, public service personnel, and non-professionals, are now engaged in the preparation and provision of metadata for Web resources. Additionally, the computer has been called on to carry a large share of the labor involved in information processing and organization. The results are often amazing and sometimes dismaying. This raises the question of how to maintain consistency and quality while struggling to achieve efficiency. The answer perhaps lies somewhere between a total reliance on human power and a complete delegation to technology.

Retrieval Models

The new landscape presented by the Web challenges established information retrieval models to provide the power to navigate networked resources with the same levels of efficiency in precision and recall achieved with traditional resources. In her deliberation of subject cataloging in the online environment, Marcia J. Bates pointed out the importance of bringing into consideration search capabilities "*Online search capabilities themselves constitute a form of indexing.* Subject access to online catalogs is thus a combination of original indexing and what we might call 'search capabilities indexing'" (Bates 1989). In contemplating the most effective subject approaches to networked resources, we need to take into account the different models currently used in information retrieval. In addition to the Boolean model, various ranking algorithms and other retrieval models are also implemented. The Boolean model, based on exact

matches, is used in most OPACs and many commercial databases. On the other hand, the vector and the probabilistic models are common on the Web, particularly in full-text analysis, indexing, and retrieval (Korfhage 1997; Salton 1994). In these models, the loss of specificity normally expected from traditional subject access tools is compensated to a certain degree by methods of statistical ranking and computational linguistics, based on term occurrences, term frequency, word proximity, and term weighting. These models do not always yield the best results but, combined with automatic methods in text processing and indexing, they have the ability to handle a large amount of data efficiently. They also give some indication of future trends and developments.

Subject Access on the Web

What kinds of subject access tools are needed in this environment? We may begin by defining their functional requirements. Subject access tools are used:

- to assist searchers in identifying the most efficient paths for resource discovery and retrieval
- to help users focus their searches
- to enable optimal recall
- to enable optimal precision
- to assist searchers in developing alternative search strategies
- to provide all of the above in the most efficient, effective, and economical manner

To fulfill these functions in the networked environment, there are certain operational requirements, the most important of these being interoperability and the ability to handle a large amount of resources efficiently. The blurred boundaries of information spaces demand that disparate systems can work together for the benefit of the users. Interoperability enables users to search among resources from a multitude of sources generated and organized according to different standards and approaches. The sheer size of the Web demands attention and presents a particularly critical challenge. For years, a pressing issue facing the libraries has been the large backlogs. If the definition of arrearage is a large number of books waiting in the backroom to be cataloged, then think of Web resources as a huge arrearage sitting in the front yard. How to impose bibliographic control on those resources of value in the most efficient and economical manner possible -- in essence achieving scalability -- is an important mission of the library and information profession. To provide users with a means to seamlessly utilize these vast resources, the operational requirements may be summarized as:

- interoperability among different systems, metadata standards, and languages
- flexibility and adaptability to different information communities, not only different types of libraries, but also other communities such as museums, archives, corporate information systems, etc.
- extensibility and scalability to accommodate the need for different degrees of depth and different subject domains
- simplicity in application, i.e., easy to use and to comprehend
- versatility, i.e., the ability to perform different functions
- amenability to computer application

In 1997, in order to investigate the issues surrounding subject access in the networked environment, ALCTS (Association of Library Collections and Technical Services) established two subcommittees: Subcommittee on Metadata and Subject Analysis and Subcommittee on Metadata and Classification. Their reports are now available (ALCTS 1999, 1999a). Some of their recommendations will be discussed later in this paper.

Verbal Subject Access

While subject access to networked resources is available, there is much room for improvement. Greater usage of controlled vocabulary may be one of the answers. During the past three decades, the introduction and increasing popularity and, in some cases, total reliance on free-text or natural language searching have brought a key question to the forefront: Is there still a need for controlled vocabulary? To information professionals who have appreciated the power of controlled vocabulary, the answer has always been a confident "yes." To others, the affirmative answer became clear only when searching began to be bogged down in the sheer size of retrieved results. Controlled vocabulary offers the benefits of consistency, accuracy, and control (Bates 1989), which are often lacking in the free-text approach. Even in the age of automatic indexing and with the ease in keyword searching, controlled vocabulary has much to offer in improving retrieval results and in alleviating the burden of synonym and homograph control placed on the user. For many years, Elaine Svenonius has argued that using controlled vocabulary retrieves more relevant records by placing the burden on the indexer rather than the user (Svenonius 1986; Svenonius 2000). Recently, David Batty makes a similar observation on the role of controlled vocabulary in the Web environment: "There is a burden of effort in information storage and retrieval that may be shifted from shoulder to shoulder, from author, to indexer, to index language designer, to searcher, to user. It may even be shared in different proportions. But it will not go away" (Batty 1998).

Controlled vocabulary most likely will not replace keyword searching, but it can be used to supplement and complement keyword searching to enhance retrieval results. The basic functions of controlled vocabulary, i.e., better recall through synonym control and term relationships and greater precision through homograph control, have not been completely supplanted by keyword searching, even with all the power a totally machine-driven system can bring to bear. To this effect, the ALCTS Subcommittee on Metadata and Subject Analysis recommends the use of a combination of keywords and controlled vocabulary in metadata records for Web resources (ALCTS 1999a).

Subject heading lists and thesauri began as catalogers' and indexers' tools, as a source of, and an aid in choosing, appropriate index terms. Later, they were also made available to users as searching aids, particularly in online systems. Traditionally, controlled vocabulary terms embedded in metadata records have been used as a means of matching the user's information needs against the document collection. Subject headings and descriptors, with their attendant equivalent and related terms, facilitate the searcher's ability to make an exact match of search terms against assigned index terms. Manual mapping of users' input terms to controlled vocabulary terms--for example, consulting a thesaurus to identify appropriate search terms--is a tedious process and has never been widely embraced by end-users. With the availability of online thesaurus-display, the mapping is greatly facilitated by allowing the user to browse

and select controlled vocabulary terms in searching. Controlled vocabulary thus serves as the bridge between the searcher's language and the author's language.

Even in free-text and full-text searching, keywords can be supplemented with terms "borrowed" from a controlled vocabulary to improve retrieval performance. Participating researchers of TREC (the Text Retrieval Conference), the large-scale cross-system search engine evaluation project, have found that "the amount of improvement in recall and precision which we could attribute to NLP [natural language processing] appeared to be related to the type and length of the initial search request. Longer, more detailed topic statements responded well to LMI [linguistically motivated indexing], while terse one-sentence search directives showed little improvement" (Strzalkowski et al. 2000). Because of the term relationships built in a controlled vocabulary, the retrieval system can be programmed to automatically expand an original search query to include equivalent terms, post-up or down to hierarchically related terms, or suggest associative terms. Users typically enter simple natural language terms (Drabenstott 2000), which may or may not match the language used by authors. When the searcher's keywords are mapped to a controlled vocabulary, the power of synonym and homograph control could be invoked and the variants of the searcher's terms could be called up (Bates 1998). Furthermore, the built-in related controlled terms could also be brought up to suggest alternative search terms and to help users focus their searches more effectively. In this sense, controlled vocabulary is used as a query-expansion device. It can be used to complement uncontrolled terms and terms from lexicons, dictionaries, gazetteers, and similar tools, which are rich in synonyms, but often lacking in relational terms. In the vector and probabilistic retrieval models, using a conflation of variant and related terms often yield better results than relying on the few "key" words entered by the searcher. Equivalent and related terms in a query provide context for each other. Including additional search terms from a controlled vocabulary can improve the ranking of retrieved items.

Classification and Subject Categorization

With regard to knowledge organization, traditionally, classification has been used in American libraries primarily as an organizational device for shelf-location and for browsing in the stacks. It has often been used also as a tool for collection management, for example, assisting in the creation of branch libraries and in the generation of discipline-specific acquisitions or holdings lists. In the OPAC, classification has regained its bibliographic function through the use of class numbers as access points to MARC records. To continue the use of class numbers as access points, the ALCTS Subcommittee on Metadata and Subject Analysis recommends that this function be extended to other types of metadata records by including in them class numbers, but not necessarily the item numbers, from existing classification schemes (ALCTS 1999a).

In addition to the access function, the role of classification has been expanded to those of subject browsing and navigational tools for retrieval on the Web. In its study of the use of classification devices for organizing metadata, the ALCTS Subcommittee on Metadata and Classification has identified seven functions of classification: location, browsing, hierarchical movement, retrieval, identification, limiting/partitioning, and profiling (ALCTS 1999).

With the rapid growth of networked resources, the enormous amount of information available on the Web cries out for organization. When subject categorization devices first became popular among Web information providers, they resembled broad classification schemes, but many were lacking the rigorous hierarchical structure and careful conceptual organization found in established schemes. Many library portals, which began with a collection of a limited number of selected electronic resources offering only keyword searching and/or an alphabetical listing, have adopted broad subject categorization schemes when the collection of electronic resources became voluminous and unwieldy (Waldhart et al. 2000). Some of these subject categorization devices are based on existing classification schemes, e.g., Internet Public Library Online Texts Collection based on the Dewey Decimal Classification (DDC) and CyberStacks(sm) based on the Library of Congress Classification (LCC) (McKiernan 2000); others represent home-made varieties.

Subject categorization defines narrower domains within which term searching can be carried out more efficiently and enables the retrieval of more relevant results. Combination of subject categorization with term searching has proven to be an effective and efficient approach in resource discovery and data mining. In this regard, classification or subject categorizing schemes function as information filters, used to efficiently exclude large segments of a database from consideration of a search query (Korfhage 1997).

Recent Research on Subject Access Systems

Before we explore the potential directions for future development of traditional subject access tools, let us also examine some of the recent research efforts and their implications for current and future methods of subject indexing and access. A huge body of research has been reported in the literature. Three areas of experimentation that I consider to have important bearings on subject access tools are automatic indexing, mapping terms and data from different sources, and integrating different subject access tools.

Automatic indexing

In the past few decades, some of the most important research in the field of information storage and retrieval has been focused on automatic indexing. Beginning with the pioneer efforts in the 1970s, various techniques, including term weighting, statistical analysis of text, and computational linguistics, have been developed and applied. More recent examples include OCLC's Scorpion project, which uses automatic methods to perform subject recognition and to generate machine-assigned DDC numbers for electronic resources (Shafer 1997). Another OCLC project, WordSmith, (Godby and Reighart 1998), applying computational linguistics to implement a series of largely statistical filters, investigates the feasibility of extracting subject terminology directly from raw text. An extension of this project, called Extended WordSmith, applies a similar technique to the automatic generation of thesaural terms. On the more practical side, the recent implementation of the LEXIS/NEXIS SmartIndexing Technology combines controlled vocabulary features with an indexing algorithm to arrive at a relevance score or percentage based on criteria such as term frequency, weight, and location in document in indexing LEXIS/NEXIS news collections (Quint 1999; Tenopir 1999).

Mapping terms and data from different sources

Mapping natural-language expressions typical of end-user search queries and of automatically extracted index terms to more structured subject-language is an area that has been explored and holds great promise (Svenonius 2000). A recent example is the "Entry Vocabulary Modules" project at the University of California-Berkeley, which explores the possibility of mapping "ordinary language queries" to indexing terms based on metadata subject vocabularies unfamiliar to the user, including classification numbers, subject headings, and descriptors from various subject- or domain-specific vocabularies (Buckland et al. 1999).

On another front, numerous efforts have focused on mapping subject data from different vocabulary sources, including free-text terms extracted from full texts, controlled vocabularies, classification data, and name authority data. Because the networked environment is open and multifarious, multiple tools for resource description and subject access are often used side-by-side. In this open environment, use of multiple controlled vocabularies within the same system is not uncommon. Harmonization of different vocabularies, similar or analogous to crosswalks among metadata schemes, is an important issue. Even before the advent of the World Wide Web, mapping subject terms from multiple thesauri was a topic of great interest and concern. An example was Carol Mandel's investigation to resolve the problems caused by using multiple vocabularies within the same online system (Mandel 1987). Much progress has been made in biomedical vocabularies. The Unified Medical Language System (UMLS) Metathesaurus currently maps biomedical terms from over fifty different biomedical vocabularies, some in multiple languages (Nelson 1999; National Library of Medicine 2000). A general metathesaurus covering all subjects is still lacking. Outside of the library context, there are also efforts to map index terms from different sources. An example is WILSONLINE's OmniFile, which results from merging index terms from six H.W. Wilson indexes into one index file.

On a broader scale, indexes from different language sources also need to be interoperable. Mapping between controlled vocabularies in different languages is an issue of great interest particularly in the international community. MACS (Multilingual ACcess to Subject), an ongoing international project involving Swiss, German, French, and British national libraries, attempts to link subject authority files in three different languages, Schlagwortnormdatei (SWD, German), RAMEAU (French), and the Library of Congress Subject Headings (English) (Landry 2000).

Mapping between subject headings and class numbers is not new. Past efforts have focused mainly on facilitating subject cataloging and indexing. Examples include the linking of many LCC numbers to headings in the *Library of Congress Subject Headings (LCSH)* list and the inclusion of abridged DDC numbers in the *Sears List of Subject Headings* (Sears). More recently, there have been efforts to map between DDC numbers and LCSH (Vizine-Goetz 1998). OCLC's WordSmith project mentioned earlier demonstrates that subject terms can be identified and extracted automatically from raw texts and mapped to existing classification schemes such as DDC (Godby and Reighart 1998). Diane Vizine-Goetz demonstrates how results from the research projects WordSmith and ExTended Concept Trees can be used together to enhance DDC (Vizine-Goetz 1997). The same techniques should be applicable to LCC also. With the implementation of the CORC (Cooperative Online Resource Catalog) project, results of many of OCLC's research projects have converged in practice. Actual application includes the automatic

generation of subject data and DDC numbers in metadata records. A most impressive feature of CORC that can yield great benefit is the capability of mapping names and subject words and phrases input by catalogers or indexers or those automatically generated from websites to entries in subject and name authority files.

Integrating different subject access tools

In the manual environment, subject headings and classification systems have more or less operated in isolation from each other. Technology offers the possibility of integrating tools of different sorts to enhance retrieval results as well as facilitate subject cataloging and indexing. The merging or integration of classification with controlled vocabulary holds great potential. Numerous research projects have been undertaken and some of the designs have been tested. For example, Karen Markey's project incorporated the Dewey Decimal Classification as a retrieval tool alongside subject searching in an online system (Markey 1986). Her research was built on AUDACIOUS, an earlier project using UDC as the index language with nuclear science literature (Freeman and Atherton 1968).

In a system called Cheshire, Ray Larson, used a method called "classification clustering," combined with probabilistic retrieval techniques, to improve subject searching in the OPAC. Starting with LC call numbers and using probabilistic ranking and weighting mechanisms, Larson demonstrates that class numbers combined with subject terms generated from titles of documents and subject headings in MARC records can enhance access points and improve greatly the retrieval results. The integration of different types of access is significant, as Larson observes: "The topical access points of the MARC records used in online catalogs, such as the classification numbers, subject headings, and title keywords, have usually been treated in strict isolation from each other in search. The classification clustering method is one way of effectively combining these difference 'clues' to the database contents" (Larson 1991).

Traditional Tools in the Networked Environment

The concepts surrounding subject access have been explored in relation to the configuration of the Web landscape and retrieval models. In this context, a question that can be raised is: How well can existing subject access tools fulfill the requirements of networked resources? More specifically, how adequate are traditional tools such as LCSH, LCC, and DDC in meeting the challenges of effective and efficient subject retrieval in the networked environment?

Library of Congress Subject Headings

With regard to LCSH specifically, a basic question is whether a new controlled vocabulary more suited to the requirements of electronic resources should be constructed. The ALCTS Subcommittee on Metadata and Subject Analysis deliberated on this question and examined the options relating to the choice of subject vocabulary in metadata records. After considering the options of developing a new vocabulary or adopting or adapting one or more existing vocabularies, the Subcommittee recommends the latter option (ALCTS 1999a). For a general controlled vocabulary covering all subjects, the Subcommittee recommends the use of LCSH or Sears with or without modifications. Among the reasons for retaining

LCSH are: (1) LCSH is a rich vocabulary covering all subject areas, easily the largest general indexing vocabulary in the English language; (2) there is synonym and homograph control; (3) it contains rich links (cross references indicating relationships) among terms; (4) it is a pre-coordinate system that ensures precision in retrieval; (5) it facilitates browsing of multiple-concept or multi-faceted subjects; and, (6) having been translated or adapted as a model for developing subject headings systems by many countries around the world, LCSH is a de facto universal controlled vocabulary. In addition, there is another major advantage. Retaining LCSH as subject data in metadata records would ensure semantic interoperability between the enormous store of MARC records and metadata records prepared according to various other standards.

While the vocabulary, or semantics, of LCSH has much to contribute to the management and retrieval of networked resources, the way it is currently applied has certain limitations: (1) because of its complex syntax and application rules, assigning LC subject headings according to current Library of Congress policies requires trained personnel; (2) subject heading strings in bibliographic or metadata records are costly to maintain; (3) LCSH, in its present form and application, is not compatible in syntax with most other controlled vocabularies; and, (4) it is not amenable to search engines outside of the OPAC environment, particularly current Web search engines. These limitations mean that applying LCSH properly in compliance with current policy and procedures entails the following requirements:

- trained catalogers and indexers
- systems with index browsing capability
- systems with online thesaurus display
- sophisticated users (Drabenstott 1999)

In the networked environment, such conditions often do not prevail. What direction and steps need to be taken for LCSH to overcome these limitations and remain useful in its traditional roles as well as to accommodate other uses? Pondering the viability of LCSH in the networked environment, the ALCTS Subcommittee on Metadata and Subject Analysis recommends separating the consideration regarding semantics from that relating to application syntax, in other words, distinguishing between the vocabulary (LCSH per se) and the indexing system (i.e., how LCSH is applied in a particular implementation).

This recommendation involves several important concepts that need to be reviewed. Semantics and syntax are two distinct aspects of a controlled vocabulary. Semantics concerns the source vocabulary, i.e., what appears in the term list (e.g., a thesaurus or a subject headings list) that contains the building blocks for constructing indexing terms or search statements. It covers the scope and depth, the selection of terms to be included, the forms of valid terms, synonym and homograph control, and the syndetic (cross-referencing) devices. Semantics should be governed by well-defined principles of vocabulary structure.

At the heart of the syntax concept is the representation of complex subjects through combination, or coordination, of terms representing different subjects or different facets (defined as families of concepts that share a common characteristic (Batty 1998)) of a subject. There are two aspects of syntax: term construction and application syntax. Term construction, i.e., how words are put together to represent concepts in the thesaurus, is an aspect of semantics and is a matter of principle; while application syntax,

i.e., how thesaurual terms are put together to reflect the contents of documents in the metadata record, is a matter of policy, determined by practical factors such as user needs, available resources, and search engines and their capabilities.

Enumeration (i.e., the listing of pre-established multiple-concept index terms in the thesaurus) and faceting (i.e., the separate listing of single-concept or single-facet terms defined in distinctive categories based on common, shared characteristics) are aspects of term construction, while precoordination and postcoordination relate to application syntax. Term combination can occur at any of three stages in the process of information storage and retrieval: (1) during vocabulary construction; (2) at the stage of cataloging or indexing; or, (3) at the point of retrieval. When words or phrases representing different subjects or different facets of a subject are pre-combined at the point of thesaurus construction, we refer to the process as enumeration. When term combination occurs at the stage of indexing or cataloging, we refer to the practice as precoordination. In contrast, postcoordination refers to the combination of terms at the point of retrieval. A totally enumerative vocabulary is by definition precoordinated. On the other hand, a faceted controlled vocabulary--i.e., a system that provides individual terms in clearly defined categories, or facets-- may be applied either precoordinately or postcoordinately. A faceted scheme hence is more flexible. An example of a rigorously faceted, precoordinate system is PRECIS (previously used in the British National Bibliography). Another example is the Universal Subject Environment (USE) system, proposed in a recent article by William E. Studwell, which contains faceted terms and uses special punctuation marks as facet indicators (Studwell 2000). On the other hand, current indexing systems used in abstracting and indexing services employing controlled vocabularies are typically postcoordinated. Whether a precoordinate approach or a postcoordinate approach is used in a particular implementation is a matter of policy and is agency-specific. In the remainder of this paper, we will focus on the semantics and term construction issues.

Because of the varied approaches to retrieval in different search environments and the different needs of diverse user communities, a vocabulary that is flexible enough to be used either precoordinately or postcoordinately would be the most viable. A faceted scheme can accommodate different application syntaxes, from the most complex (e.g., full-string approach typically found in OPACs) to the simplest (descriptor-like terms used in most indexes) and would also allow different degrees of sophistication. The advantages of a faceted controlled vocabulary can be summarized as follows:

- simple in structure
- flexible in application (i.e., able to accommodate a tiered approach to allow different levels of subject representation)
- amenable to software applications (Batty 1998)
- amenable to computer-assisted indexing and validation
- interoperable with the majority of modern indexing vocabularies
- easier and more economical to maintain than an enumerated vocabulary

On the last point regarding efficient thesaurus maintenance, Batty remarks: "Facet procedure has many advantages. By organizing the terms into smaller, related groups, each group of terms can be examined more easily and efficiently for consistency, order, hierarchical relationships, relationships to other groups,

and the acceptability of the language used in the terms. The faceted approach is also useful for its flexibility in dealing with the addition of new terms and new relationships. Because each facet can stand alone, changes can usually be made easily in a facet at any time without disturbing the rest of the thesaurus" (Batty 1998). Thus, a faceted LCSH will be easier to maintain. With the current LCSH, updating terminology sometimes can be a tedious operation. For example, when the heading "Moving-pictures" was replaced in 1987 by "Motion pictures," approximately 400 authority records were affected! (El-Hoshy 1998).

A faceted LCSH is by no means a new idea. Earlier advocates of such an approach include Pauline A. Cochrane (1986) and Mary Dykstra (1988). To remain viable in the networked environment, a controlled vocabulary, such as LCSH, must be able to accommodate different retrieval models mentioned earlier as well as different application policies. Outside of the OPAC, most search engines, including many used in library portals for Web resources, lack the ability to accommodate full-string browsing and searching. Even among systems that can handle full strings, their capabilities and degrees of sophistication also vary. With a faceted vocabulary, it will not be an either/or proposition between the precoordinate full-string application and the postcoordinate approach, but rather a question of how LCSH can be made to accommodate both and any variations in between, thus ensuring maximum flexibility and scalability in terms of application. Mechanisms for full-string implementation of LCSH are already in place; for example, in the OPAC environment, with highly trained personnel and the searching and browsing capabilities of integrated systems, the full-string syntax has long been employed in creating subject headings in MARC records. In the heterogeneous environment outside of the OPAC, we need a more flexible system in order to accommodate different applications. LCSH can become such a tool, and its use can be extended to various metadata standards and with different encoding schemes. Investigations and experiments on the viability of LCSH have already begun. Using LCSH as the source vocabulary, FAST (Faceted Application of Subject Terminology), a current OCLC research project, explores the possibility and feasibility of a postcoordinate approach by separating time, space, and form data from the subject heading string (Chan et al. in press).

Now we come to the question of where LCSH stands currently in becoming a viable system for the networked environment. LCSH began in the late nineteenth century as an enumerative scheme. It gradually took on some of the features of a faceted system, particularly in the adoption of commonly used form subdivisions and the increasing use of geographic subdivisions. In the latter part of the twentieth century LCSH has taken further steps, ever so cautiously, in the direction of more rigorous faceting. In 1974, the Library of Congress took a giant leap forward in expanding the application of commonly used subdivisions by designating a large number of frequently used topical and form subdivisions "free-floating," thus allowing great flexibility in application. The adoption of BT, NT, RT in the 11th (1988) edition rendered LCSH more in line with thesaural practice. After the Subject Subdivisions Conference held in 1991 (The Future of Subdivisions, 1992), the Library of Congress has embarked on a program to convert many of the topical subdivisions into topical main headings. Finally, in 1999, the implementation of subfield $v for form subdivisions in the 6xx (subject-related) fields in the MARC format, marking the distinction between form and topical subdivisions, moved LCSH yet another step closer to becoming a faceted system. Considering the gradual steps the Library of Congress has taken over the years, even a person not familiar with the history of LCSH must conclude logically that LCSH is heading in the

direction of becoming a fully faceted vocabulary. It is not there yet; but, with further effort, LCSH can become a versatile system that is capable of functioning in heterogeneous environments and can serve as the unified basis for supporting diversified uses while maintaining semantic interoperability among them.

A faceted LCSH has a number of potential uses in the areas of thesaurus development and management, indexing, and retrieval. As mentioned earlier, to enhance the interoperability of a multitude of controlled vocabularies, a general metathesaurus covering all subjects would be most desirable (ALCTS 1999a). It will not be a trivial task, but the first question the library and information profession must agree upon is whether it is something worth pursuing. LCSH, with its rich vocabulary--the largest in the English language--can serve as a basis or core of such a metathesaurus.

From a different perspective, LCSH could also be used as the basis for generating subject- or discipline-specific controlled vocabularies or special-purpose thesauri. The *AC Subject Headings* (formerly *Subject Headings for Children's Literature*) sets a precedent. Other examples include a large "superthesaurus" proposed by Bates (1989), with a rich entry vocabulary as a part of a friendly front-end user interface for the OPAC. While many subject domains and disciplines such as engineering, art, and biomedical sciences have their own controlled vocabularies, many specialized areas and non-library institutions still lack them. These include for-profit as well as non-profit organizations, government agencies, historical societies, special-purpose museums, consulting firms, fashion design companies, to name a few. Many of these rely on their curators or researchers, most of whom have not been trained in bibliographic control, to take responsibility for organizing Internet resources. Having a comprehensive subject access vocabulary to draw and build upon would be of tremendous help in developing their specialized thesauri.

To move LCSH further along the way towards becoming a faceted vocabulary, if indeed such is the direction to be followed, more can be done to its semantics. Aspects of particular concern that need close scrutiny and re-thinking include principles of term selection, enhanced entry vocabulary, rigorous term relationships, and particularly term construction.

Library of Congress Classification (LCC) and Dewey Decimal Classification (DDC)

In recent years, with the support of the OCLC Research Office, DDC has made great strides in adapting to the networked environment and becoming a useful tool for organizing electronic resources. For example, the newly developed WebDewey contains, in addition to the DDC/LCSH mapping feature first developed in Dewey for Windows, an automated classification tool for generating candidate DDC numbers during metadata record creation. It has taken LCC somewhat longer because its voluminous schedules have only recently been converted to the MARC format. Let us hope that the Library Congress can now turn its attention to making LCC a useful tool not only in the library stacks but also as an organizing tool of networked resources. Results and insights gained from experimental and actual implementations of Web application of DDC and other classification schemes should be applicable to LCC as well.

Existing classification schemes have already been adopted or adapted to a limited extent for use as subject categorization devices for Web resources. Examples include the adaptation of DDC in NetFirst and CyberDewey and the use of LCC outlines in Cyberstacks. In this particular role, existing classification

schemes need greater flexibility and more attention to their structure. Adaptability of classification schemes can take the form of flexibility in the depth of hierarchy and variability in the collocation of items in the array. The requirement of depth varies from application to application. As a tool for shelf-location and bibliographic arrangement, considerable depth in classification is required, as evidenced in the growth of both DDC and LCC in the past. As a navigating tool typified by the subject categorization schemes used in the popular Web directories, broad schemes are often sufficient. What is needed is a flexibility of depth and the amenability to the creation of classificatory structures focused on specific subject domains. Flexibility in depth has always been a feature of DDC and UDC, with the availability of abridged, medium, and full versions, in recognition of the different needs of school, public, and research libraries. LCC has not yet demonstrated this flexibility. This is an area worth exploring.

The principle of literary warrant, i.e., basing the development of a scheme or system on the nature and extent of resources being described and organized, operates in the Web environment as well as in the print environment. In the development of subject categorization schemes used in popular Web information services, such as Yahoo! and Northern Light (Ward 1999) as well as many library portals, we have often witnessed the gradual extension from simple, skeletal outlines to increasingly elaborate structures--almost a mirror of the development of classification schemes in the early days. Flexibility in the collocation of topics in an array would also be helpful, if the same topics in an array could be arranged or re-arranged in different orders depending on the target audiences. For example, the categorization scheme in NetFirst uses the DDC structure, but modifies the arrangement of the categories to suit its target users (Vizine-Goetz 1997).

Observing recent uses of classification-like structures on the Web and the tortuous re-inventing and re-discovering of classification principles in both research and practice (Soergel 1999a), one sees a need for both broad/general (covering all subjects) and close/detailed (subject- or domain-specific) classification schemes. Portals found on websites of general libraries, ranging from school and public libraries to large academic libraries that cover a broad range of subject domains, need schemes of varying depths with a top-down approach, beginning with the broadest level and moving down to narrower subjects as needed. On the other hand, portals that serve special clientele often need specialized schemes with more details. These often require a bottom-up approach starting with topics identified from a collection of documents focusing on a specific theme or mission. How to organize these topics into a coherent structure has often stymied those not trained in the principles and techniques of knowledge organization. The library and information profession can make a contribution here. Subject taxonomy schemes built around specific disciplines (art, education, human environmental sciences, mathematics, engineering), industries (petroleum, manufacturing, entertainment), consumer-oriented topics (automobiles, travel, sports), and problems (environment, aging, juvenile delinquency) can serve diverse user communities, from special libraries to corporate or industry information centers to personal resource collections.

For domain- and subject-specific organizing schemes I suggest a modular approach. In building special-purpose thesauri mentioned earlier, LCSH could serve as the source vocabulary, and DDC or LCC could be used to facilitate the identification and extraction of terms related to specific subjects or domains and could provide the underlying hierarchical structure. Where more details are needed in a particular scheme, terms can be added to the basic structure as needed, thus making the specialized scheme an extension of

the main structure and vocabulary. Developing these modules with a view of fitting them as nodes, even on a very broad level, into the overall classification structures of meta-schemes such as DDC and LCC can go a long way to ensure their future interoperability.

As mentioned earlier, the merging or integration of controlled subject vocabulary with classification in order to facilitate both information storage and retrieval has great potential, because they complement each other. A subject heading or descriptor represents a particular topic treated from all perspectives, while classification gathers related topics viewed from the same perspective. Traditionally, each performs its specific function and contributes to information organization and retrieval more or less in isolation. Together, they have the potential of improving efficiency as well as effectiveness. Schemes simple and logical in design lend themselves to interoperate efficiently with each other. How to combine the salient features of a rich vocabulary like LCSH and the structured hierarchy found in classification schemes such as LCC and DDC to improve retrieval of networked resources remains a fertile field for research and exploration.

Conclusion

The sheer volume of available networked resources demands efficiency in knowledge management. Of course, we intend to provide quality and to maintain consistency also. Content representation schemes and systems design must meet halfway--a combination of the intellect and technology, capitalizing on the power of the human mind and the capabilities of the machine. Technology has provided an impetus in the creation of an enormous amount of information; it can also help in its effective and efficient management and retrieval (Soergel 1999). A proper balance in the distribution of efforts between human intellect and technology can ensure both quality and efficiency in helping users gain the maximum benefits from the rich resources that are available in the networked environment. Already, technology has helped create many useful devices for efficient management and application of traditional tools, for example, Dewey for Windows, the WebDewey, and ClassificationPlus. These developments are encouraging. In the near future, we may expect also new applications which can help us not only do the same things better and more efficiently, but also maximize the power of existing subject access tools hitherto not yet exploited.

References

ALCTS/CCS/SAC/Subcommittee on Metadata and Classification. (1999). Final Report. http://www.ala.org/alcts/organization/ccs/sac/metaclassfinal.pdf

ALCTS/CCS/SAC/Subcommittee on Metadata and Subject Analysis. (1999a). Subject Data in the Metadata Record: Recommendations and Rationale: A Report from the ALCTS/CCS/SAC/Subcommittee on Metadata and Subject Analysis. http://www.ala.org/alcts/organization/ccs/sac/metarept2.html

Bates, Marcia J. (1998). Indexing and Access for Digital Libraries and the Internet: Human, Database,

and Domain Factors. *Journal of the American Society for Information Science*. 49(13):1185-1205.

Bates, Marcia J. (October 1999). The Invisible Substrate of Information Science. *Journal of the American Society for Information Science*. 50(12):1043-1050.

Bates, Marcia J. (October 1989). Rethinking Subject Cataloging in the Online Environment. *Library Resources & Technical Services*. 33(4):400-412.

Bates, Marcia J. (1986). Subject Access in Online Catalogs: A Design Model. *Journal of the American Society for Information Science*. 37:357-76.

Batty, David. (November 1998). WWW -- Wealth, Weariness or Waste: Controlled Vocabulary and Thesauri in Support of Online Information Access. *D-Lib Magazine*. http://www.dlib.org/dlib/november98/11batty.html

Buckland, Michael, et al.. (January 1999). Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies. *D-Lib Magazine*. http://www.dlib.org/dlib/january99/buckland/01buckland.html

Burton, Paul F. (Summer 1998). Issues and Challenges of Subject Access. *Catalogue & Index*. 128:1-7.

Chan, Lois Mai, Eric Childress, Rebecca Dean, Edward T. O'Neill, and Diane Vizine-Goetz. (in press). A Faceted Approach to Subject Data in the Dublin Core Metadata Record. *Journal of Internet Cataloging*.

Chan, Lois Mai. (1995). *Library of Congress Subject Headings: Principles and Application*. 3rd ed. Englewood, CO: Libraries Unlimited.

Cochrane, Pauline A. (1986). *Improving LCSH for Use in Online Catalogs: Exercises for Self-Help with a Selection of Background Readings*. Littleton, CO: Libraries Unlimited.

Drabenstott, Karen M. (2000). Web Search Strategies. In *Saving the User's Time through Subject Access Innovation*, edited by William J. Wheeler. Champaign, IL: Graduate School of Library and Information Science, University of Illinois.

Drabenstott, Karen M., Schelle Simcox, and Marie Williams. (Summer 1999). Do Librarians Understand the Subject Headings in Library Catalogs? *Reference & User Services Quarterly*. 38(4):369-87.

Dykstra, Mary. (March 1 1988). LC Subject Headings Disguised as a Thesaurus. *Library Journal*. 113:42-46.

El-Hoshy, Lynn M. (August 1998). Charting a Changing Language with LCSH. *Library of Congress Information Bulletin*. 57(8):201.

Freeman, Robert R. and Atherton, Pauline. (April 1968). *AUDACIOUS - An Experiment with an On-Line, Interactive Reference Retrieval System Using the Universal Decimal Classification as the Index Language in the Field of Nuclear Science.* New York, NY: American Inst. of Physics, (QPX02169).

The Future of Subdivisions in the Library of Congress Subject Headings System: Report from the Subject Subdivisions Conference, sponsored by the Library of Congress, May 9-12, 1991, ed. Martha O'Hara Conway. Washington, DC: Library of Congress, Cataloging Distribution Service, 1992.

Godby, C. J. (1998). The Wordsmith Toolkit. *Annual Review of OCLC Research 1997.* Available at http://www.oclc.org/oclc/research/publications/review97/godby/godby_wordsmith.htm

Godby, C. Jean and Reighart, Ray R. (1998). The WordSmith Indexing System. http://www.oclc.org/oclc/research/publications/review98/godby_reighart/wordsmith.htm

Godby, C. J. and R. Reighart. (1998a). The Wordsmith Project Bridges the Gap between Tokenizing and Indexing. *OCLC Newsletter*, July 1998. Available at http://www.oclc.org/oclc/new/n234/rsch_wordsmith_research_project.htm.

Korfhage, Robert R. (1997). The Matching Process. In Information Storage and Retrieval. New York: Wiley. (pp. 79-104).

Landry, Patrice. (2000). The MACS Project: Multilingual Access to Subjects (LCSH, RAMEAU, SWD). Classification and Indexing Workshop, 66th IFLA Council and General Conference, Meeting No. 181. http://www.ifla.org/IV/ifla66/papers/165-181e.pdf

Larson, Ray R. (1991). Classification Clustering, Probabilistic Information Retrieval, and the Online Catalog. *The Library Quarterly* 61(2):133-173.

Mandel, Carol A. (1987). *Multiple Thesauri in Online Library Bibliographic Systems: A Report Prepared for Library of Congress Processing Services.* Washington, DC: Library of Congress, Cataloging Distribution Service.

Markey, Karen and Anh N. Demeyer. (1986). *Dewey Decimal Classification Online Project: Evaluation of a Library Schedule and Index Integrated into the Subject Searching Capabilities of an Online Catalog: Final Report to the Council on Library Resources.* Dublin, OH: OCLC.

Moen, William E. (March, 2000). Interoperability for Information Access: Technical Standards and Policy Considerations. *Journal of Academic Librarianship*. 26(2):129-32.

National Library of Medicine. (February 2000). Fact Sheet: UMLS (r) Metathesaurus (r) http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

Nelson, Stuart J. (1999). The Role of the Unified Medical Language System (UMLS) in Vocabulary Control: CENDI Conference "Controlled Vocabulary and the Internet." http://www.dtic.mil/cendi/pres_arc.html

Olson, Tony and Gary Strawn. (March 1997). Mapping the LCSH and MeSH Systems. *Information Technology and Libraries*. 16(1):5-19.

Quint, Barbara. (December 1999). Company Dossier Product Emerges from LEXIS-NEXIS SmartIndexing Technology. *Information Today*. 16(1):18-19.

Salton, Gerard. (February 1994). Automatic Structuring and Retrieval of Large Text Files. *Communications of the ACM*. 37:97-108.

Salton, Gerard. (1991). Developments in Automatic Text Retrieval. *Science*. 253:974-980.

Shafer, Keith. (October/November 1997). Scorpion Helps Catalog the Web. *Bulletin of the American Society for Information Science*. 24(1):28-29.

Shafer, Keith E. (1996). Automatic Subject Assignment Via the Scorpion System. *Annual Review of OCLC Research*, pp. 20-21.

Soergel, Dagobert. (September 1999). Enriched Thesauri as Networked Knowledge Bases for People and Machines: Paper Presented at the CENDI Conference Controlled Vocabulary and the Internet, Bathesda, MD, 1999 September 29. http://www.dtic.mil/cendi/presentations/cendisoergel.pdf

Soergel, Dagobert. (1999a). The Rise of Ontologies or the Reinvention of Classification. *Journal of the American Society for information Science*. 50(12):1119-1120.

Strzalkowski, Tomek et al.. (2000). Natural Language Information Retrieval: TREC-8 Report. http://trec.nist.gov/pubs/trec8/papers/ge8adhoc2.pdf

Studwell, William E. (2000). USE, the Universal Subject Environment: A New Subject Access Approach in the Time of the Internet. *Journal of Internet Cataloging*. 2(3/4):197-209.

Svenonius, Elaine. (2000). *The Intellectual Foundation of Information Organization*. Cambridge, MA: MIT Press.

Svenonius, Elaine. (1986). Unanswered Questions in the Design of Controlled Vocabularies. *Journal of the American Society for Information Science*. 37:331-40.

Tenopir, Carol. (November 1, 1999). Human or Automated, Indexing Is Important. *Library Journal*.

124(18):34,38.

Vizine-Goetz, Diane. (1996). Classification Research at OCLC. *Annual Review of OCLC Research*, pp. 27-33.

Vizine-Goetz, Diane. (October/November 1997). From Book Classification to Knowledge Organization: Improving Internet Resource Description and Discovery. *Bulletin of the American Society for Information Science.* 24(1):24-27.

Vizine-Goetz, Diane. (May/June 1998). Subject Headings for Everyone: Popular Library of Congress Subject Headings with Dewey Numbers. *OCLC Newsletter*. 233:29-33.

Waldhart, Thomas J., Joseph B. Miller, and Lois Mai Chan. (March 2000). Provision of Local Assisted Access to Selected Internet Information Resources by ARL Academic Libraries. *Journal of Academic Librarianship*. 26(2):100-109.

Ward, Joyce. (1999). Indexing and Classification at Northern Light: Presentation to CENDI Conference "Controlled Vocabulary and the Internet," Sept 29, 1999. http://www.dtic.mil/cendi/pres_arc.html

Younger, Jennifer A. (Winter 1997). Resource Description in the Digital Age. *Library Trends*. 45(3):462-87.

Library of Congress
December 19, 2000
Comments: lcweb@loc.gov

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Lois Mai Chan
Professor
School of Library and Information Science
University of Kentucky
Lexington, KY 40506-0039

Exploiting LCSH, LCC, and DDC to Retrieve Networked Resources

About the presenter:

Lois Mai Chan is the author of numerous articles and books on cataloging, classification, subject indexing and online retrieval, including *Library of Congress Subject Headings: Principles and Application*; *Cataloging and Classification: An Introduction*; and *Immroth's Guide to the Library of Congress Classification*. She co-authored *Dewey Decimal Classification: A Practical Guide and Thesauri Used in Online Databases*. From 1986 to 1991, she served as the chair of the Decimal Classification Editorial Policy Committee. Her research interests include classification, controlled vocabulary, authority control, metadata, and retrieval of Web resources. In 1989, she was awarded the Margaret Mann Citation for Outstanding Achievement in Cataloging and Classification given by ALA. In 1992, she received the Distinguished Service Award from the Chinese-America n Librarians Association. In 1999, she was chosen for the Best of LRTS Award for the Best Article Published in 1998.

Full text of paper is available

Summary:

Vocabulary control for improved precision and recall and structured

BEST COPY AVAILABLE

Cataloging Directorate Home Page

Library of Congress Home Page

organization for efficient shelf location and browsing have contributed to effective subject access to library materials. The question is whether existing tools can continue to function satisfactorily in dealing with web resources. In our effort to identify library resource description needs and future directions, the online public access catalog (OPAC) should be viewed as a part of the overall information storage and retrieval apparatus on the web rather than something apart from it. Deliberations on the future of bibliographic control and the tools used for its implementation should take into consideration the nature of the web, the characteristics of web resources, and the variety of information retrieval approaches and mechanisms now available and used on the web. Operational conditions on the web are often less structured than in the OPAC environment. While traditional subject access tools such as subject headings and classification schemes have served library users long and well, there are certain limitations to their extended applicability to networked resources. These include the need of trained catalogers for their proper application according to current policies and procedures, the cost of maintenance, and their incompatibility with most tools now used on the web. To meet the challenges of web resources, certain operational requirements must be taken into consideration, the most important being the ability to handle a large volume of resources efficiently and interoperability across different information environments and among a variety of retrieval models. Schemes that are scalable in semantics and flexible in syntax, structure, and application are more likely to be capable of meeting the requirements of a diversity of information retrieval environments and the needs of different user communities. Library of Congress Subject Headings (LCSH), the Library of Congress Classification (LCC), and the Dewey Decimal Classification (DDC) have long been the main staples of subject access tools in library catalogs. Recent deliberations of the Association of Library Collections and Technical Services (ALCTS) Subcommittee on Subject Analysis and Metadata and research findings suggest that in order to extend their usefulness as subject access tools in the web environment, traditional schemes must undergo rigorous scrutiny and re-thinking, particularly in terms of their structure and the way they are applied. Experimentation conducted on subject access schemes in surrogate-based WebPACs and metadata-processed systems demonstrate the potential benefit of structured approaches to description and organization of web resources. Research findings indicate that sophisticated technology can be used to extend the usefulness and to enhance the power of traditional tools. Together, they can provide approaches to content retrieval that may offer improved or perhaps even better subject access than many methods currently used in full-text document analysis and retrieval on the web.